Supporting Information

Boosting medical diagnostics by pooling independent judgments

Ralf H.J.M. Kurvers, Stefan M. Herzog, Ralph Hertwig, Jens Krause, Patricia A. Carney, Andy Bogart, Giuseppe Argenziano, Iris Zalaudek & Max Wolf

PART I: SUPPLEMENTARY FIGURES

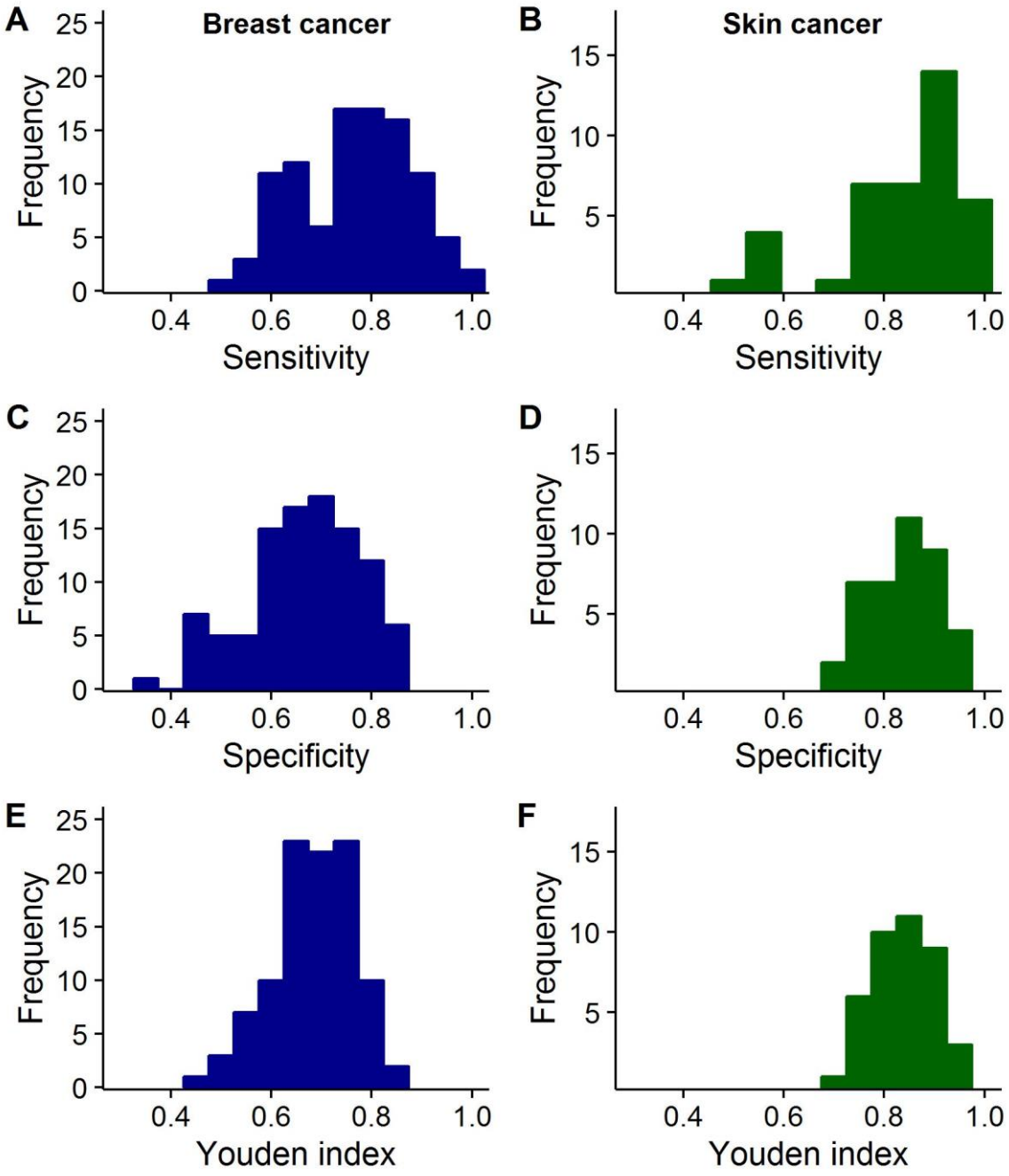

Fig. S1: **Histograms of average individual performance.**

The frequency of average individual (A, B) sensitivity, (C, D) specificity and (E, F) Youden's index of the (A, C, E) radiologists (n = 101) and (B, D, F) dermatologists (n = 40).

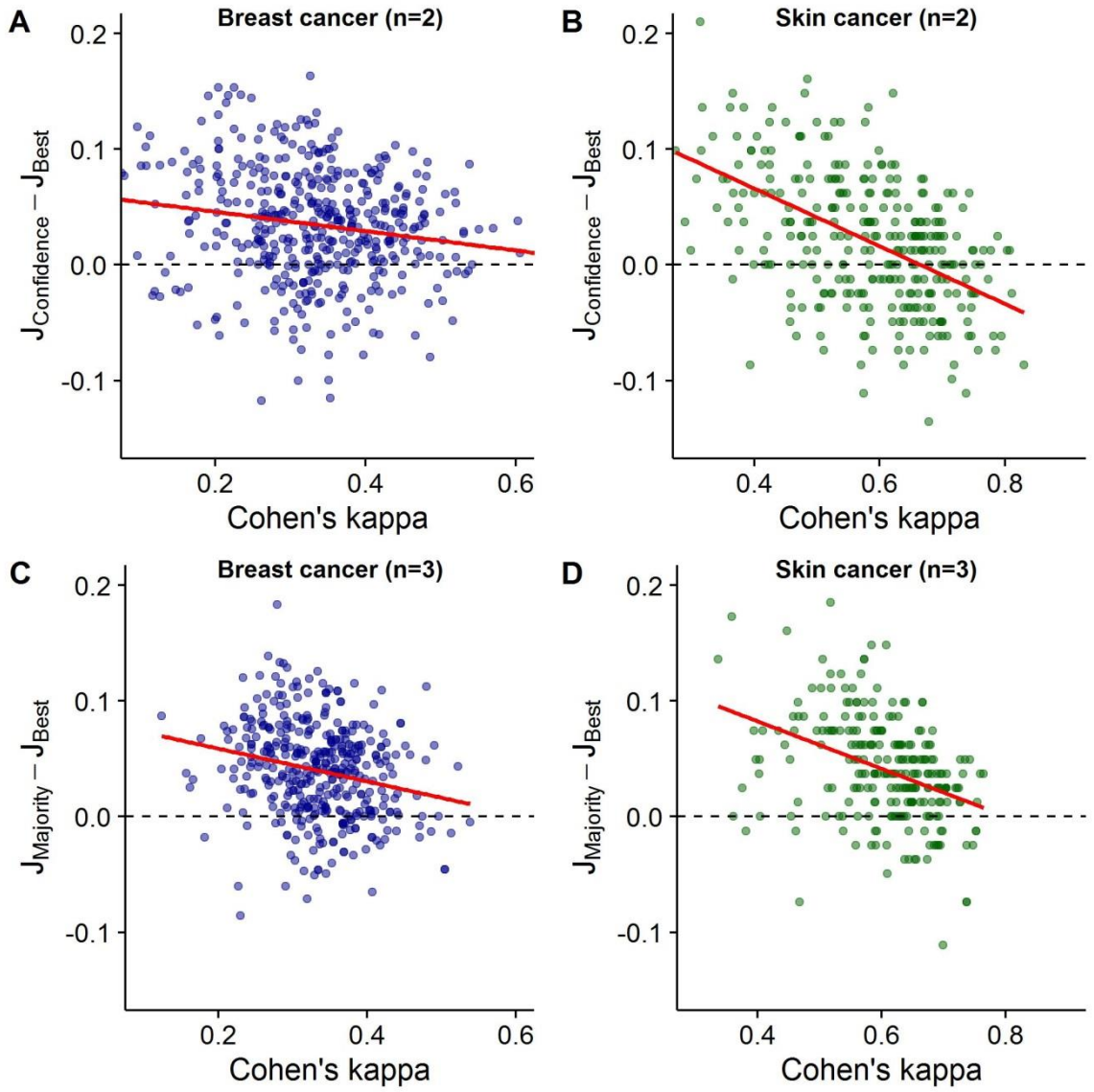

Fig. S2: **Performance of the confidence/majority rule relative to the best diagnostician in that group as a function of the independence of judgments.**

(A-D) Each dot represents a unique combination of (A, B) two or (C, D) three diagnosticians. Values above zero indicate that the confidence/majority rule outperformed the best individual in that group. Values below zero indicate that the best diagnostician outperformed the confidence/majority rule. Red lines are linear regression lines. With increasing Cohen's kappa (i.e. lower independence of judgments), the ability of groups to outperform its best member decreases. Only groups with diagnosticians of similar ability in terms of Youden's index (i.e. $\Delta J < 0.1$) are shown since this is the region where collective intelligence arises.

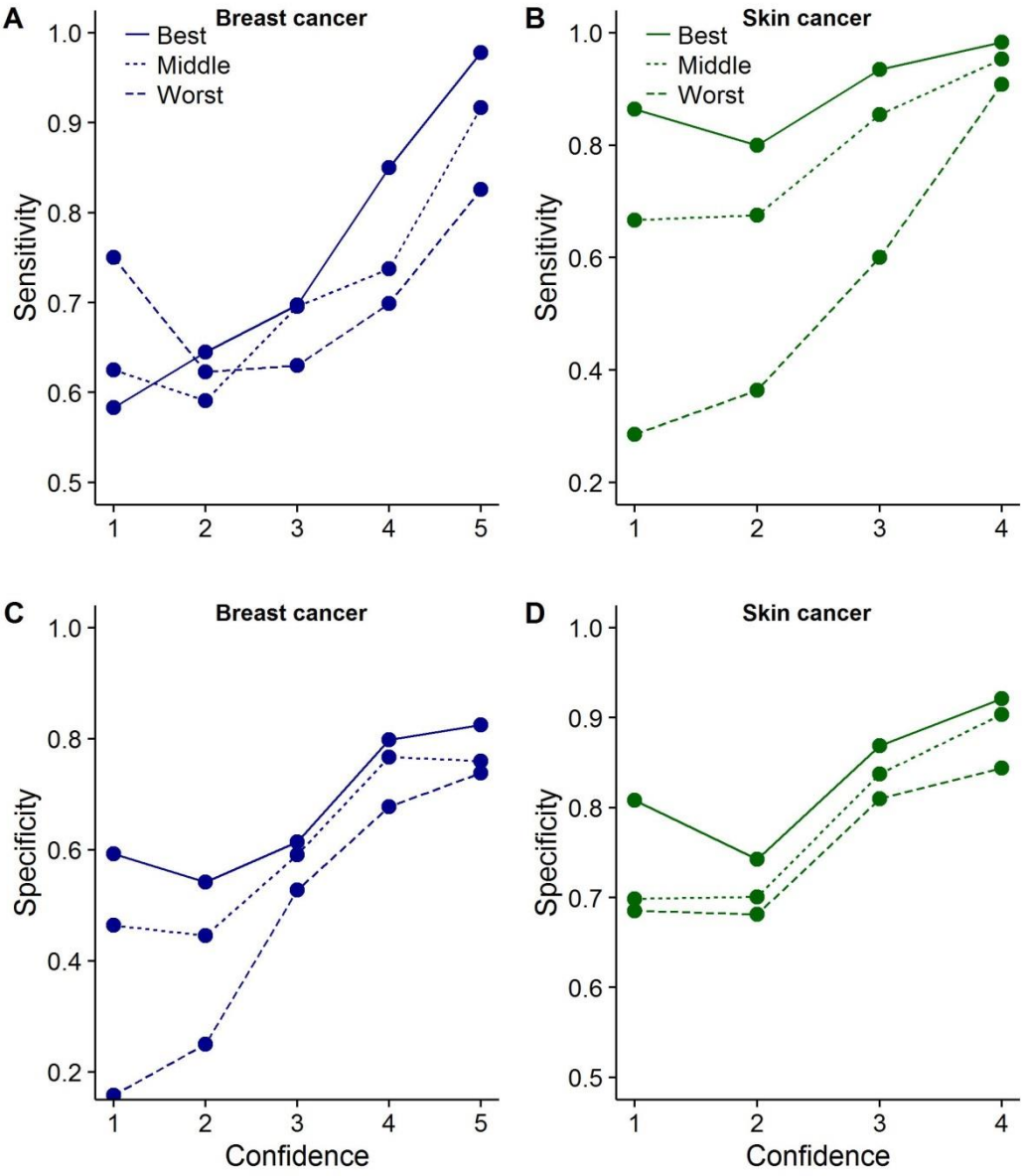

Fig. S3: Relationship between the reported confidence level and the sensitivity/specificity.

There was a positive relationship between confidence and sensitivity/specificity for the (i) best- (ii) midlevel-, and (iii) poorest performing diagnosticians (based on the Youden's index). For a given confidence level, the performance of the best diagnosticians was generally higher than the performance of the middle diagnosticians which, in turn, was generally higher than the performance of the poorest diagnosticians.

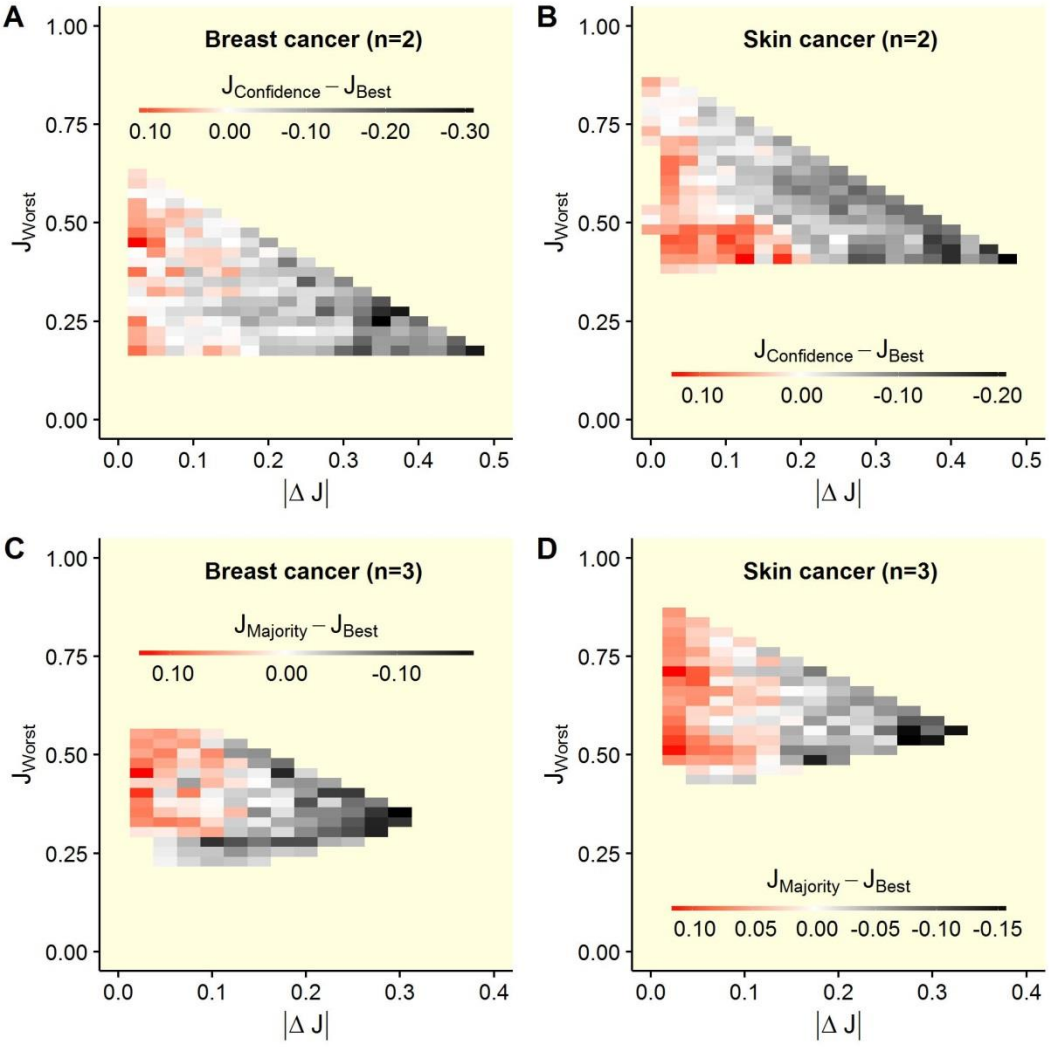

Fig. S4: Difference in performance between the confidence/majority rule and the best diagnostician in the group as a function of the performance of the poorest diagnostician.

Difference in performance between the collective intelligence rule and the best group member as a function of the difference in accuracy (Youden's index *J)* between group members (x-axis) and the accuracy of the worst group member (y-axis). Shown are results for groups of (A, B) two diagnosticians using the confidence rule, and (C, D) three diagnosticians using the majority rule. Red areas indicate that the confidence/majority rule outperformed the best diagnostician; white areas indicate no difference; grey and black areas indicate that the best diagnostician outperformed the confidence/majority rule. The confidence/majority rule outperformed the best diagnostician only when the diagnosticians' accuracy levels were similar (i.e., left part of the heat plots). This effect was present both in groups in which the worst diagnostician performed relatively well and in groups in which he/she performed poorly.

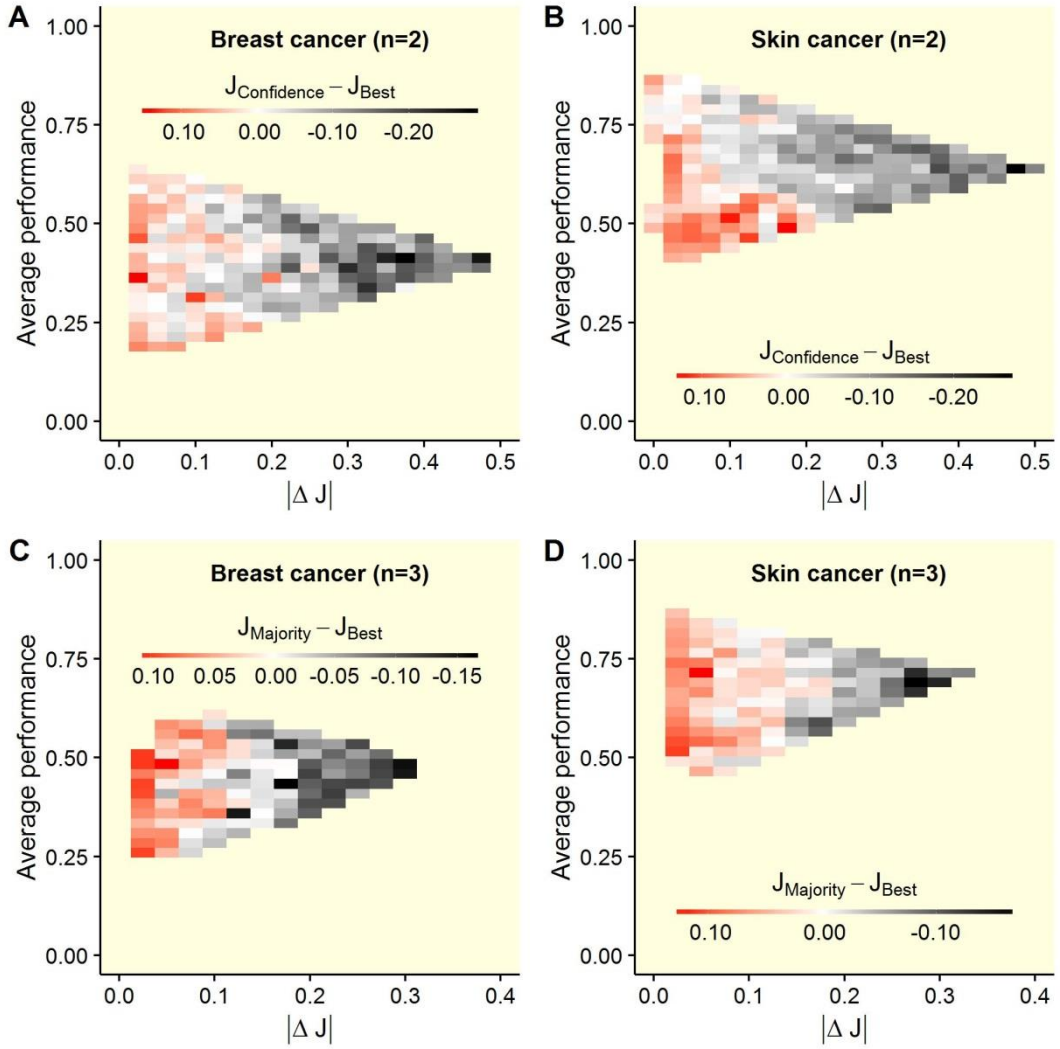

Fig. S5: Difference in performance between the confidence/majority rule and the best diagnostician in the group as a function of average individual performance.

Difference in performance between the collective intelligence rule and the best group member as a function of the difference in accuracy (Youden's index, *J)* between group members (x-axis) and the average individual performance of group members (y-axis). Shown are results for groups of (A, B) two diagnosticians using the confidence rule and (C, D) three diagnosticians using the majority rule. Red areas indicate that the confidence/majority rule outperformed the best diagnostician; white areas indicate no difference; grey and black areas indicate that the best diagnostician outperformed the confidence/majority rule. The confidence/majority rule outperformed the best diagnostician only when the diagnosticians' accuracy levels were similar (i.e., left part of the heat plots). This effect was present both in groups in which the average individual ability of diagnosticians was high and in groups in which average individual ability of diagnosticians was low.

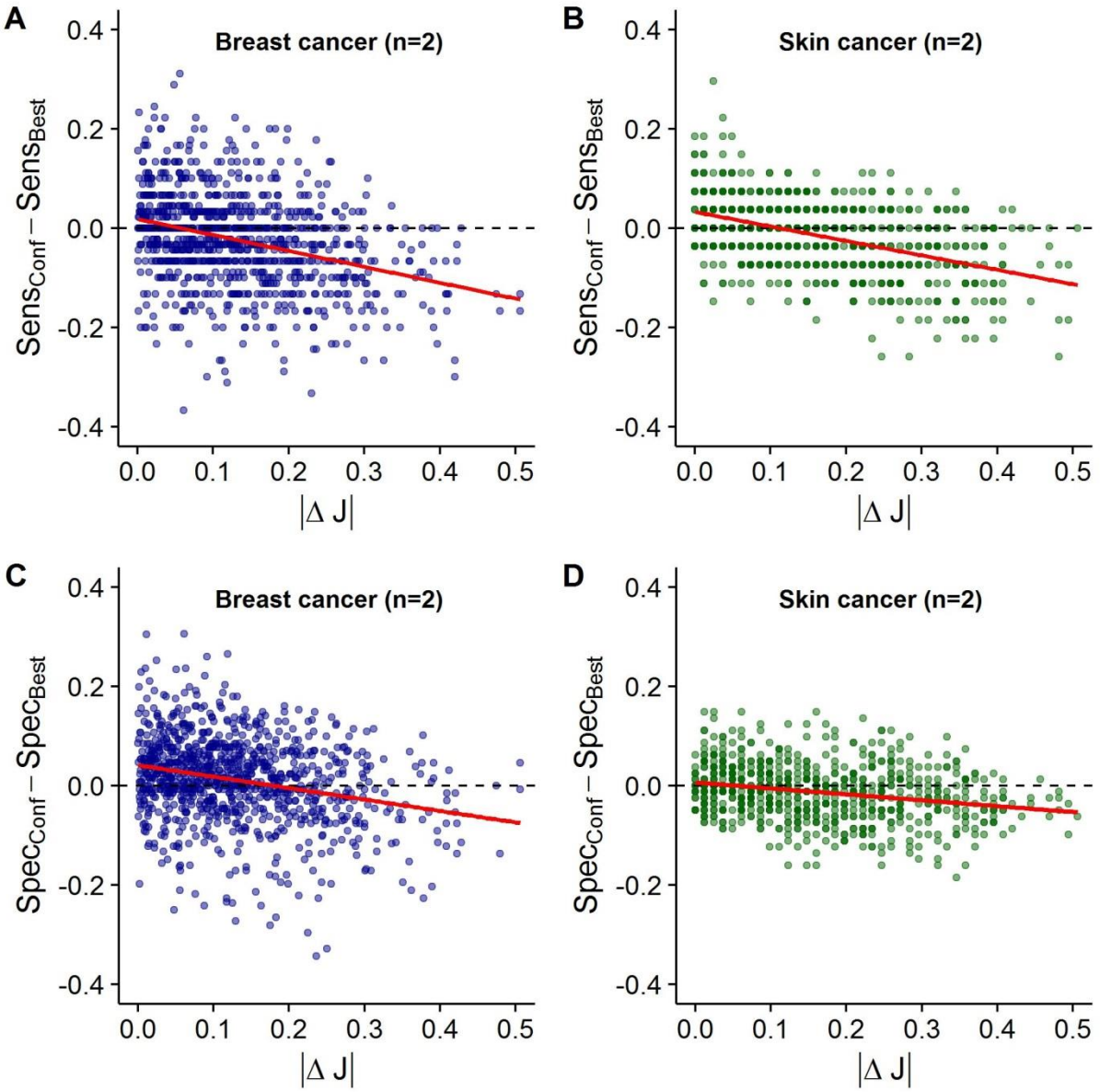

Fig. S6: Sensitivity and specificity of the confidence rule in groups of two diagnosticians relative to the best diagnostician in that group.

(A-D) Each dot represents a unique combination of two diagnosticians. Values above zero indicate that the confidence rule outperformed the best diagnostician in that group; values below zero indicate that the best diagnostician outperformed the confidence rule. Red lines are linear regression lines. In both diagnostic contexts, the confidence rule achieved higher (A, B) sensitivity and (C, D) specificity than the best diagnostician only when the diagnosticians' accuracy levels were similar.

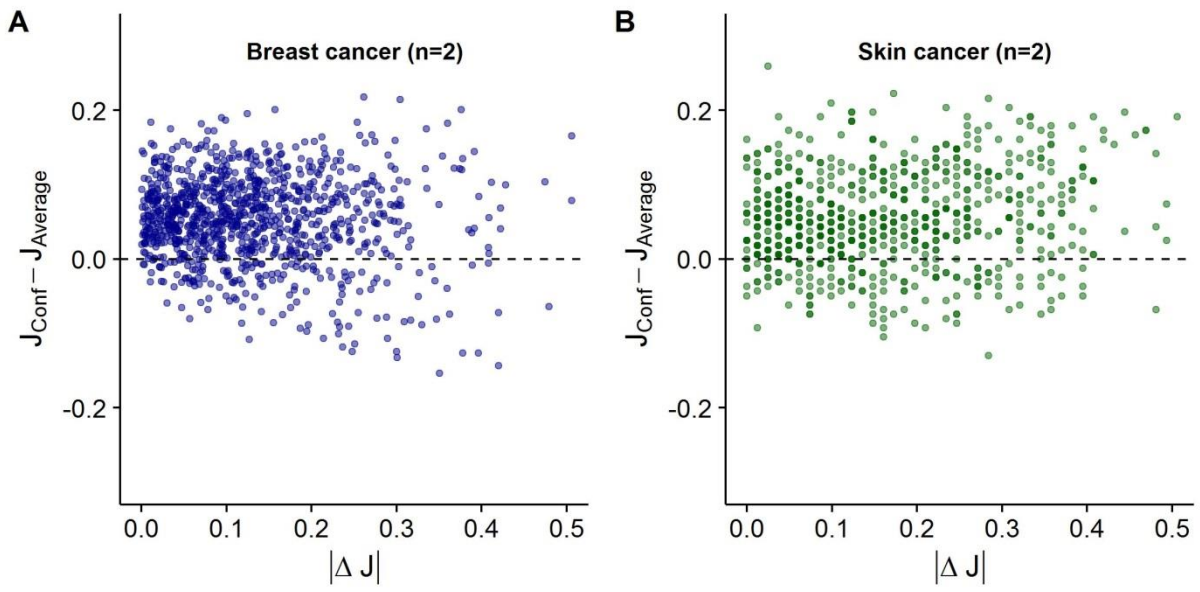

Fig. S7: **Performance of the confidence rule in groups of two diagnosticians relative to average individual performance in that group.**

(A, B) Each dot represents a unique combination of two diagnosticians. Values above zero indicate that the confidence rule outperformed the average individual in that group; values below zero indicate that the average individual outperformed the confidence rule. In both (A) breast and (B) skin cancer diagnostics, the confidence rule generally outperformed average individual performance (proportion of groups in which the confidence rule outperformed the average performance: breast cancer: 0.84; skin cancer: 0.80).

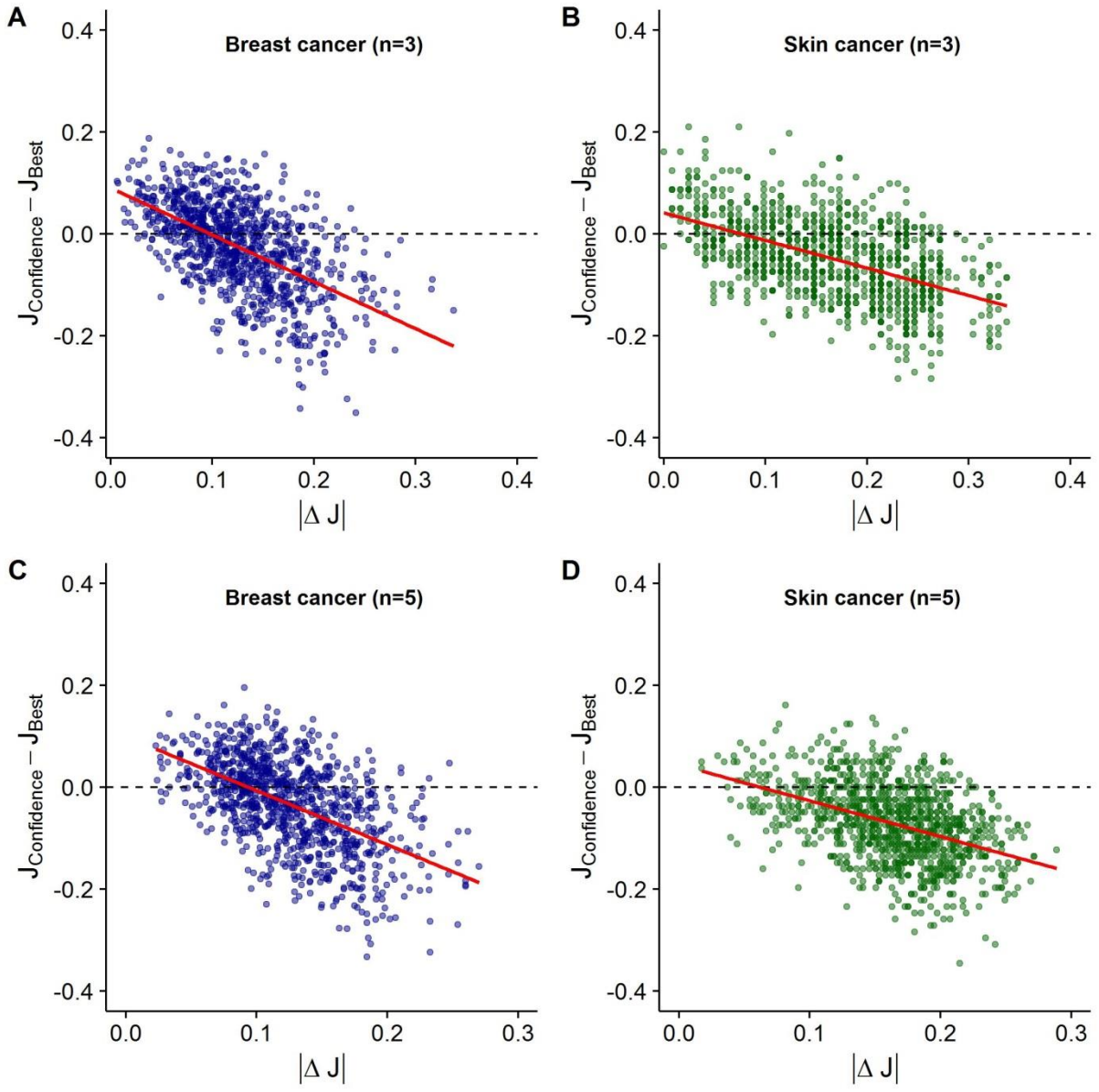

Fig. S8: Performance of the confidence rule in groups of three and five diagnosticians relative to the best diagnostician in that group.

(A-D) Each dot represents a unique combination of (A, B) three or (C, D) five diagnosticians. Values above zero indicate that the confidence rule outperformed the best diagnostician in that group. Values below zero indicate that the best diagnostician outperformed the confidence rule. Red lines are linear regression lines. In both diagnostic contexts and for both group sizes, the confidence rule outperformed the best diagnostician only when diagnosticians' accuracy levels were relatively similar.

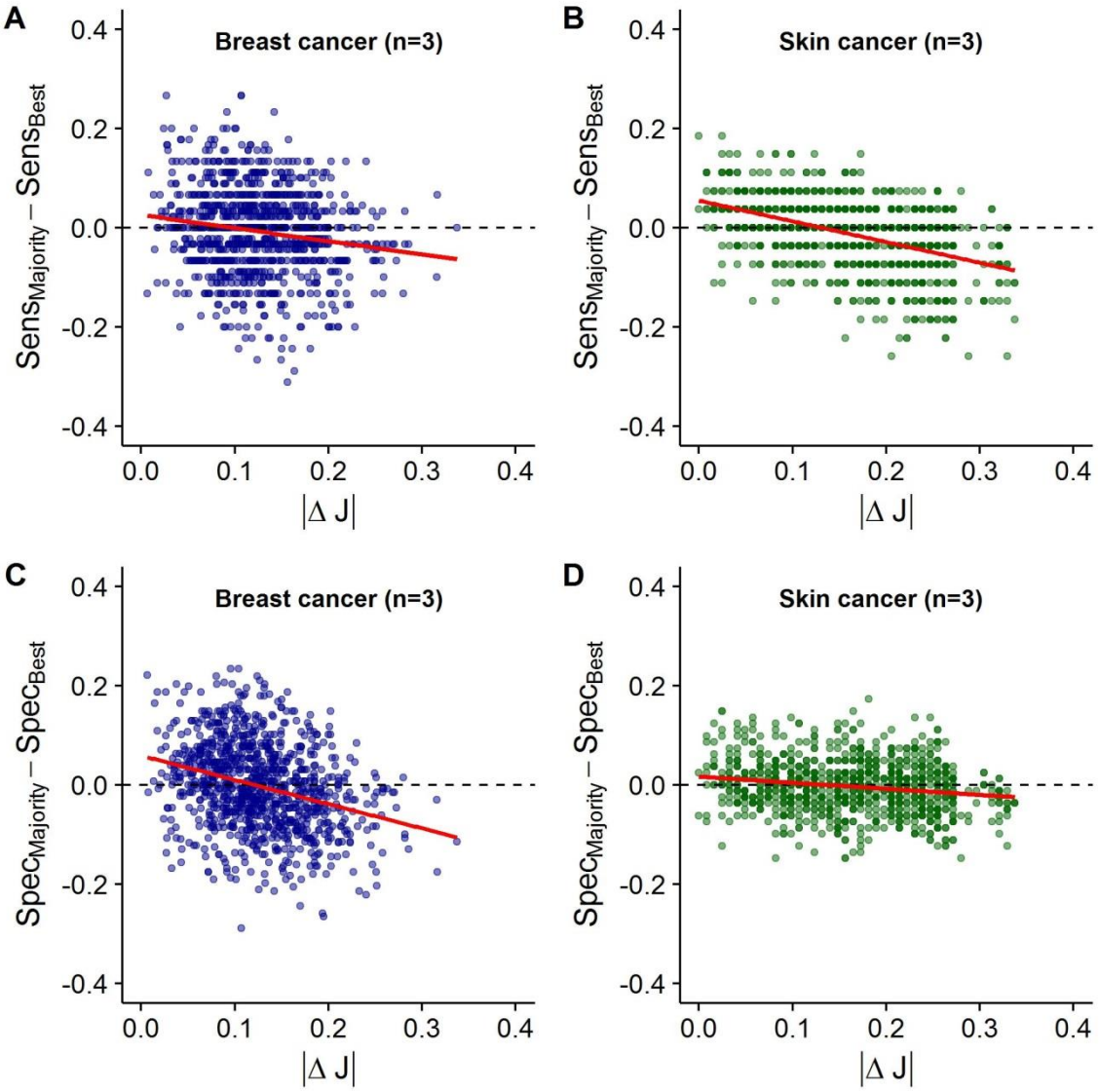

Fig. S9: Sensitivity and specificity of the majority rule in groups of three diagnosticians relative to the best diagnostician in that group.

(A-D) Each dot represents a unique combination of three diagnosticians. Values above zero indicate that the majority rule outperformed the best diagnostician in that group. Values below zero indicate that the best diagnostician outperformed the majority rule. Red lines are linear regression lines. In both diagnostic contexts, the majority rule achieved higher (A, B) sensitivity and (C, D) specificity than the best diagnostician only when diagnosticians' accuracy levels were relatively similar.

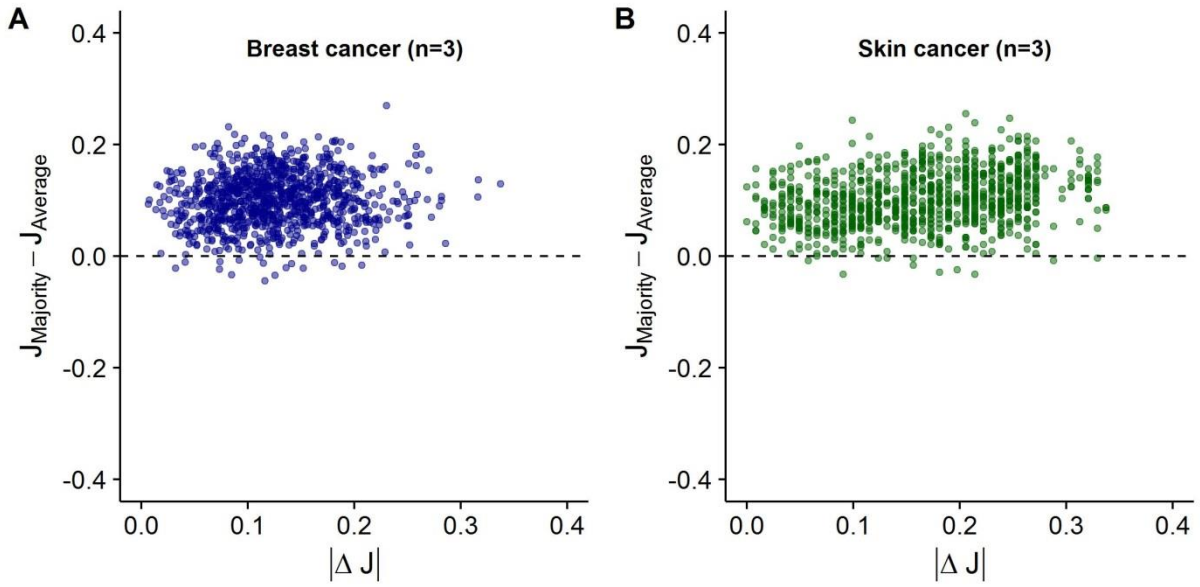

Fig. S10: **Performance of the majority rule in groups of three diagnosticians relative to average individual performance in that group.**

(A, B) Each dot represents a unique combination of three diagnosticians. Values above zero indicate that the majority rule outperformed the average individual in that group. Values below zero indicate that the average individual outperformed the majority rule. In both (A) breast and (B) skin cancer diagnostics, the majority rule generally outperformed average individual performance (proportion of groups in which the majority rule outperformed the average performance: breast cancer: 0.98; skin cancer: 0.99).

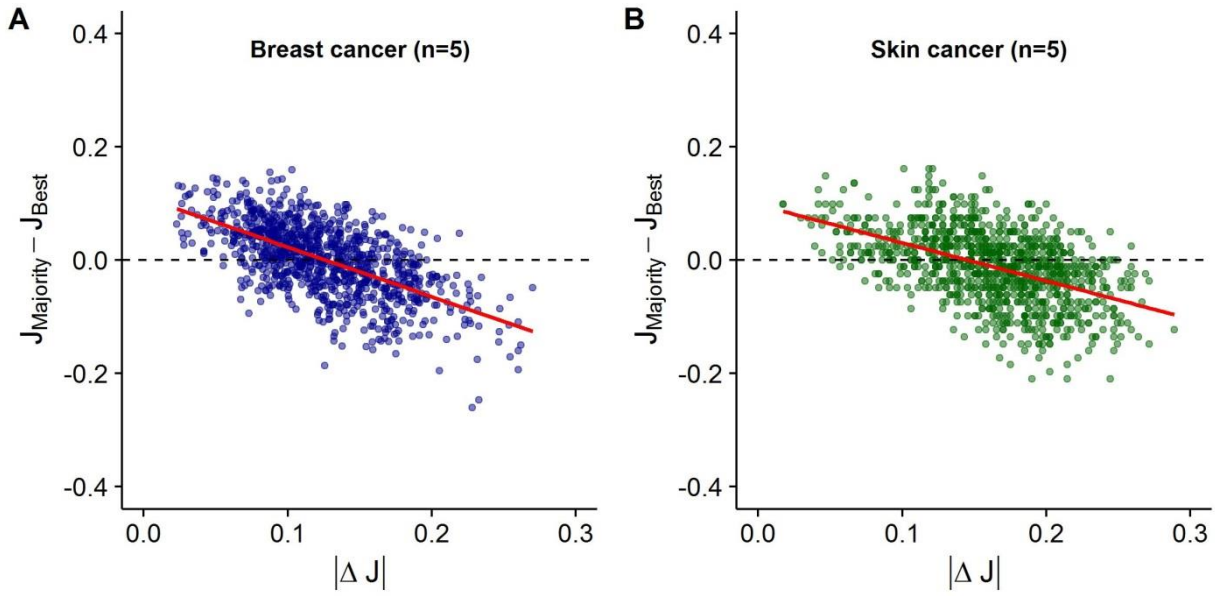

Fig. S11: Performance of the majority rule in groups of five diagnosticians relative to the best diagnostician in that group.

(A, B) Each dot represents a unique combination of five diagnosticians. Values above zero indicate that the majority rule outperformed the best individual in that group. Values below zero indicate that the best diagnostician outperformed the majority rule. Red lines are linear regression lines. In both diagnostic contexts, the majority rule outperformed the best diagnostician only when diagnosticians' accuracy levels were relatively similar.

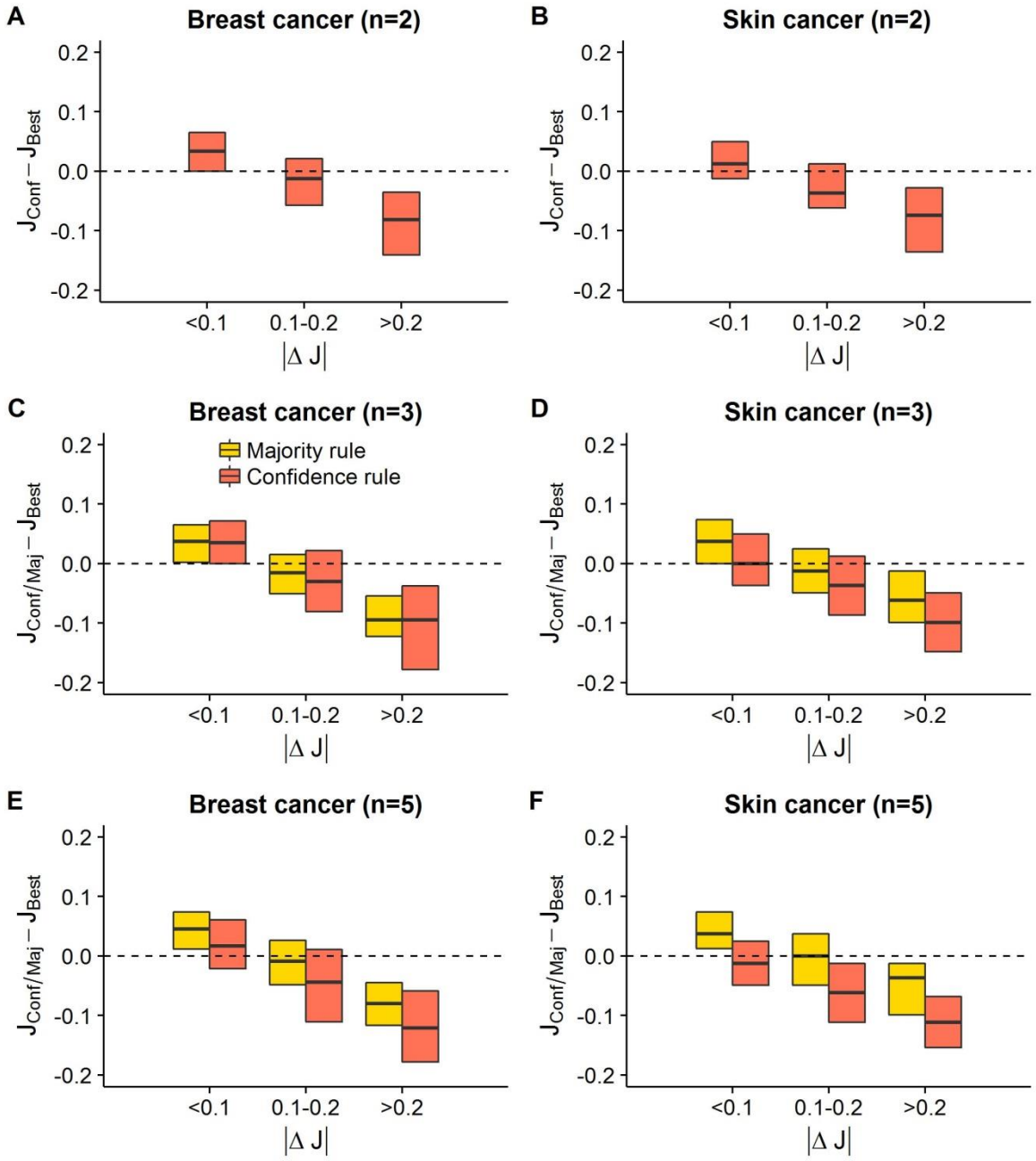

Fig. S12: Comparing the performance of the confidence and majority rule across different group sizes.

Difference in performance between the confidence/majority rule and the best group member, grouped for three categories of similarity in performance (ΔJ), for group size (A, B) two, (C, D) three and (E, F) five. For all group sizes, collective intelligence rules should only be used when diagnosticians have similar performance level (i.e., ΔJ<0.1). If so, then at group size two the confidence rule performs best, at group size three and beyond the majority rule performs best.

PART II: MODELLING ANALYSIS

In order to further understand the mechanisms underlying our results and to investigate the generality of our findings, we developed simplified analytical models of the basic scenarios investigated in our data sets. In particular, we developed models for the scenarios where two diagnosticians employ the confidence rule (Model 1) and three diagnosticians employ the majority rule (Model 2). As can be seen in the following, the results of the analytical models are fully in line with the findings of the empirical analysis.

MODEL 1: Two diagnosticians employing the confidence rule

We consider two diagnosticians with probabilities of making a correct decision p_1 and p_2, respectively. Without loss of generality, we can assume that diagnostician 1 is the more accurate of the two diagnosticians, that is $p_1 > p_2$.

We are interested in the following question. Assuming the dyad adopts the confidence rule, how does the degree of similarity in the two diagnosticians' accuracy levels affect the dyad's ability to outperform the better diagnostician?

In order to study this question, we make the simplifying assumption that, for any of the two diagnosticians, the probability of being correct, conditional on the other diagnostician being correct, corresponds to the unconditional probability of being correct, that is:

$$p_i\big|_{\text{rater j correct}} = p_i, \quad i, j = 1, 2 \text{ and } i \neq j. \tag{1}$$

The better diagnostician is outperformed whenever the probability that either (i) both diagnosticians are correct or (ii) one of the diagnosticians is correct and simultaneously has

the higher confidence score is higher than the probability that the better diagnostician is correct. That is, the better diagnostician is outperformed whenever:

$$p_1 \cdot p_2 + \left(p_1 \cdot (1 - p_2) + (1 - p_1) \cdot p_2 \right) \cdot \alpha \quad > \quad p_1, \tag{2}$$

where α corresponds to the probability that – in the case of a disagreement – the individual with the higher confidence score is correct.

This condition can be simplified to:

$$\underbrace{(1 - p_1) \cdot p_2 \cdot \alpha}_{\substack{\text{probability that an incorrect decision} \\ \text{by the better rater is improved by the poorer rater}}} - \underbrace{p_1 \cdot (1 - p_2) \cdot (1 - \alpha)}_{\substack{\text{probability that a correct decision} \\ \text{by the better rater is worsened by the poorer rater}}} > 0. \tag{3}$$

Since we are interested how the similarity in accuracy between diagnostician 1 and 2 affects the ability of the dyad to outperform the better diagnostician (while keeping the average ability in the dyad constant), we now rewrite p_1 and p_2 as:

$$\begin{aligned} p_1 &= \bar{p} + \delta, \\ p_2 &= \bar{p} - \delta, \end{aligned} \tag{4}$$

where $\bar{p}$ corresponds to the average ability of the two diagnosticians (i.e. $\frac{p_1 + p_2}{2}$) and δ is a measure of similarity between the two diagnostician (i.e. as δ increases, the similarity between the two diagnosticians decreases). We now substitute (4) into (3):

$$\underbrace{\left(1 - \left(\bar{p} + \delta\right)\right) \cdot \left(\bar{p} - \delta\right) \cdot \alpha}_{\substack{\text{probability that an incorrect decision} \\ \text{by the better rater is improved by the poorer rater}}} - \underbrace{\left(\bar{p} + \delta\right) \cdot \left(1 - \left(\bar{p} - \delta\right)\right) \cdot (1 - \alpha)}_{\substack{\text{probability that a correct decision} \\ \text{by the better rater is worsened by the poorer rater}}} > 0. \tag{5}$$

Results Model 1

Result 1.1: For diagnosticians with identical performance, i.e. $\delta = 0$, the confidence rule allows dyads to outperform any of the diagnosticians whenever the probability α with which – in case of disagreement – the individual with the higher confidence is correct is larger than 0.5, that is whenever:

$$\alpha > \frac{1}{2}. \tag{6}$$

Proof. Substitute $\delta = 0$ in (5) and solve for α.

Result 1.2: As the similarity between the two diagnosticians decreases (i.e. δ increases), the probability with which the dyad outperforms the better diagnostician decreases. More technically, the derivative of the left hand side of (5) with respect to δ is strictly negative. We note that, because of (4), this result applies to groups of differing similarity but identical average ability. Result 1.2 is illustrated in Fig. S13 below.

Proof. Taking the derivative of the left hand side of (5) with respect to δ results in

$$-(\bar{p}-\delta)\cdot\alpha-\big(1-(\bar{p}+\delta)\big)\cdot\alpha-\big(1-(\bar{p}-\delta)\big)\cdot(1-\alpha)-(\bar{p}+\delta)\cdot(1-\alpha), \tag{7}$$

which, because $0 \le \bar{p}+\delta \le 1$, $0 \le \bar{p}-\delta \le 1$ and $0 \le \alpha \le 1$, is strictly negative. This establishes Result 1.2.

Inspecting the two main terms on the left hand side of (5), we can also get a good intuition for the mechanism underlying Result 1.2. As the similarity between the two diagnosticians decreases (i.e. δ increases) two regularities simultaneously hold:

(i) The probability that the poorer diagnostician overrules an incorrect decision by the better diagnostician *decreases* because the better makes fewer incorrect decisions and the poorer makes fewer correct decisions.

(ii) The probability that the poorer diagnostician overrules a correct decision by the better diagnostician *increases* because the better makes more correct decisions and the poorer makes more incorrect decisions.

Result 1.3: Consider scenarios in which the probability that, in case of a disagreement, the individual with the higher confidence score is correct is larger than 0.5 and smaller than 1, that is, $0.5 < \alpha < 1$. In these scenarios, for low levels of similarity (i.e., high δ), the dyad performs worse than the better diagnostician; conversely, for high levels of similarity (i.e. low δ), the dyad outperforms the better diagnostician. More specifically, for any given average performance level $\bar{p}$ of the two diagnosticians in the dyad, and any level α $(0.5 < \alpha < 1)$, there exists a threshold level of similarity δ^* with the feature that dyads with a lower similarity (i.e. $\delta > \delta^*$) are outperformed by the better individual while dyads with a higher similarity (i.e. $\delta < \delta^*$) outperform the better individual. Result 1.3 is illustrated in Fig. S13 below.

Proof. From Result 1.1 we know that – whenever $\alpha > 0.5$ – the left hand side of (5) is positive for $\delta = 0$. Moreover, from Result 1.2 we know that the left hand side of (5) strictly decreases in δ. To establish Result 1.3 it is thus sufficient to show that for any particular combination of $\bar{p}$ and α $(0.5 < \alpha < 1)$, there exists a δ that turns the left hand side of (5) negative.

For this purpose, let us assume that

$$\delta = 1 - \bar{p}. \tag{8}$$

Note that this restricts our analysis to scenarios with $\bar{p} > 0.5$. Substituting (8) into (5) and rearranging the left hand side leaves us with

$$-\left(2 - 2 \cdot \bar{p}\right) \cdot (1 - \alpha), \tag{9}$$

which, because $0 < \bar{p} < 1$ and $0.5 < \alpha < 1$, is always negative. This establishes Result 1.3.

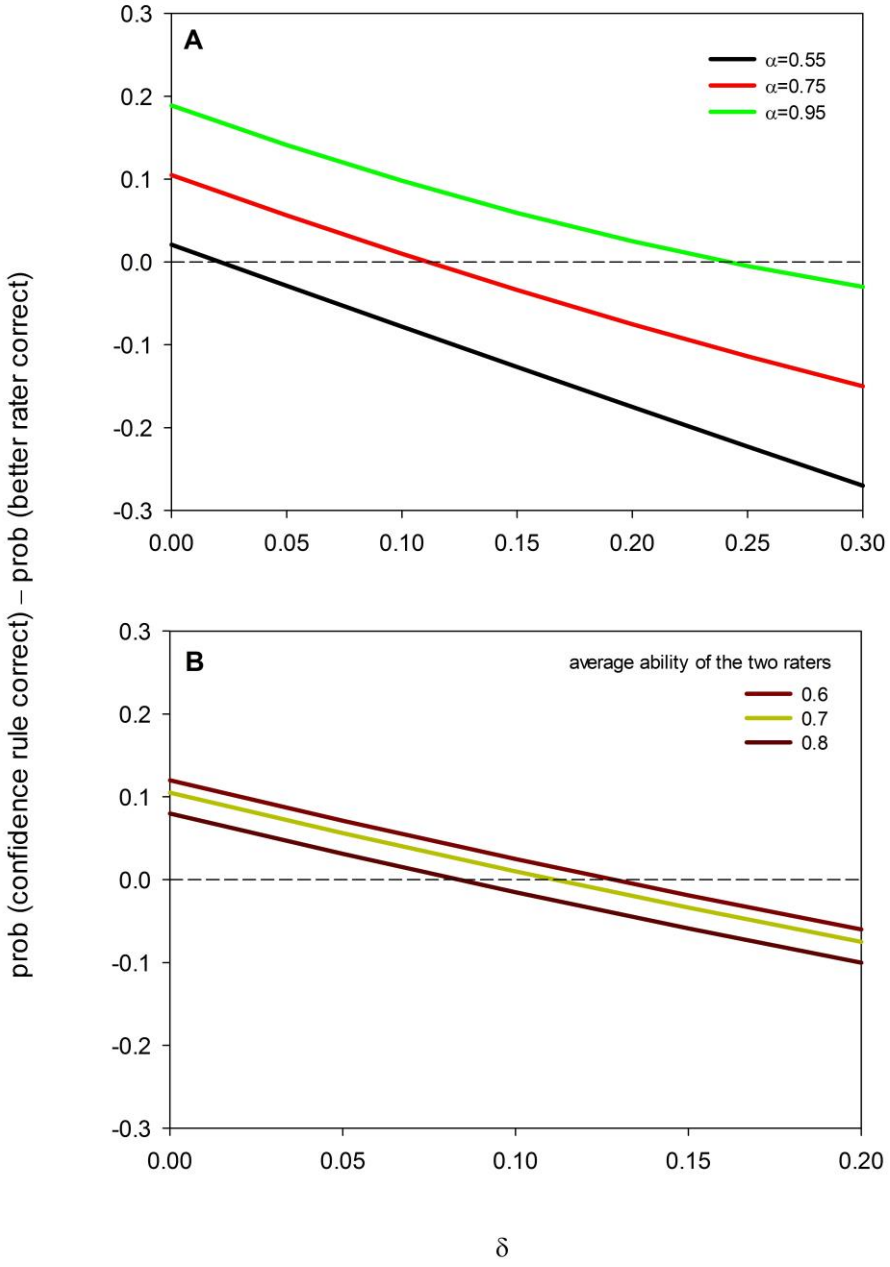

Fig. S13. As the similarity between two diagnosticians decreases (i.e. δ increases), the probability with which the dyad (when adopting the confidence rule) outperforms the better diagnostician decreases. (A) Illustrates this effect for three different levels of α (i.e. probabilities that – in case of a disagreement – the individual with the higher confidence score is correct) and a dyad with an average ability $\bar{p} = 0.7$. (B) Illustrates this effect for dyads with three different average abilities $\bar{p}$ and $\alpha = 0.75$.

MODEL 2: Three diagnosticians employing the majority rule

We consider three diagnosticians with probabilities of making a correct decision p_1, p_2 and p_3, respectively. Without loss of generality we can assume that diagnostician 1 is the most accurate of the three diagnosticians, that is $p_1 > p_2$ and $p_1 > p_3$.

We are interested in the following question. How does the degree of similarity in the three diagnosticians' accuracy levels affect the ability of that group, assuming they adopt the majority rule, to outperform the best diagnostician?

In order to analyze this question, we make two simplifying assumptions:

(i) For any of the three diagnosticians, the probability of being correct, conditional on any other diagnostician being correct corresponds to the unconditional probability of being correct, that is:

$$p_i \Big|_{\text{rater j correct}} = p_i, \quad i, j = 1,2,3 \text{ and } i \neq j. \tag{10}$$

(ii) The two poorer diagnosticians are identical in performance, that is $p_2 = p_3$.

Since we are interested how the similarity between the diagnosticians affects the ability of the three diagnosticians to outperform the best diagnostician (while keeping the average ability in the group constant), we now rewrite p_1, p_2, and p_3 as:

$$\begin{aligned} p_1 &= \bar{p} + \delta, \\ p_2 = p_3 &= \bar{p} - \frac{1}{2} \cdot \delta, \end{aligned} \tag{11}$$

where $\bar{p}$ corresponds to the average ability of three diagnosticians (i.e. $\frac{p_1+p_2+p_3}{3}$) and δ is a measure of similarity between the diagnosticians (i.e. as δ increases, the similarity between the best diagnostician and the poorer diagnosticians decreases).

Under the majority rule, the best diagnostician is outperformed whenever the probability that at least two diagnosticians are correct is higher than the probability that the best diagnostician is correct, that is the best diagnostician is outperformed whenever:

$$p_1 \cdot p_2 \cdot p_3 + p_1 \cdot p_2 \cdot (1-p_3) + p_1 \cdot (1-p_2) \cdot p_3 + (1-p_1) \cdot p_2 \cdot p_3 > p_1. \qquad (12)$$

This condition can be simplified to:

$$\underbrace{(1-p_1) \cdot p_2 \cdot p_3}_{\substack{\text{probability that an incorrect decision} \\ \text{by the best rater is improved by the two poorer raters}}} - \underbrace{p_1 \cdot (1-p_2) \cdot (1-p_3)}_{\substack{\text{probability that a correct decision} \\ \text{by the best rater is worsened by the two poorer raters}}} > 0. \qquad (13)$$

We now substitute (11) into (13):

$$\underbrace{(1-(\bar{p}+\delta)) \cdot \left(\bar{p} - \frac{1}{2} \cdot \delta\right)^2}_{\substack{\text{probability that an incorrect decision} \\ \text{by the best rater is improved by the two poorer raters}}} - \underbrace{(\bar{p}+\delta) \cdot \left(1 - \left(\bar{p} - \frac{1}{2} \cdot \delta\right)\right)^2}_{\substack{\text{probability that a correct decision} \\ \text{by the best rater is worsened by the two poorer raters}}} > 0 \qquad (14)$$

Results Model 2

Result 2.1: In case of diagnosticians with identical performance, i.e. $\delta = 0$, condition (14) reduces to the Condorcet condition, that is, it is fulfilled whenever $\bar{p} > 0.5$.

Proof. Substitute $\delta = 0$ in (14) and solve for *p*.

Result 2.2: As the similarity between diagnosticians decreases (i.e. δ increases), the probability with which the group outperforms the best diagnostician within that group decreases. More technically, the derivative of the left hand side of (14) with respect to δ is strictly negative, for all $\bar{p}$ and δ. Result 2.2 is illustrated in Fig. S14 below.

Proof. Taking the derivative of the left hand side of (14) with respect to δ results in:

$$-\left(\bar{p}-\frac{1}{2}\cdot\delta\right)^2-\left(1-\left(\bar{p}+\delta\right)\right)\cdot\left(\bar{p}-\frac{1}{2}\cdot\delta\right)-\left(1-\left(\bar{p}-\frac{1}{2}\cdot\delta\right)\right)^2-\left(\bar{p}+\delta\right)\cdot\left(1-\left(\bar{p}-\frac{1}{2}\cdot\delta\right)\right), \quad (15)$$

which – since $0\le\bar{p}+\delta\le 1$ and $0\le\bar{p}-\frac{1}{2}\cdot\delta\le 1$ –is strictly negative. This establishes Result 2.2.

Inspecting the two main terms on the left hand side of (14), we can also get a good intuition for this result. As the similarity between diagnosticians decreases (i.e. δ increases), the following two regularities simultaneously hold:

(i) The probability that the two poorer diagnosticians overrule an incorrect decision by the best diagnostician *decreases* because the best makes fewer incorrect decisions and the poorer make fewer correct decisions.

(ii) The probability that the two poorer diagnosticians overrule a correct decision by the best diagnostician *increases* because the best makes more correct decisions and the poorer make more incorrect decisions.

Result 2.3: Consider scenarios with $\bar{p}>0.5$. For low levels of similarity (i.e. high δ), the group performs worse than the best diagnostician; conversely, for high levels of similarity (i.e. low δ), the group outperforms the best diagnostician. More specifically, for any given average performance level $\bar{p}$ in the group, there exists a threshold level of similarity δ^* with the feature that groups with a lower similarity (i.e., $\delta>\delta^*$) are outperformed by the best individual while groups with a higher similarity (i.e., $\delta<\delta^*$) outperform the best individual. Result 2.2 is illustrated in Fig. S14 below.

Proof. From Result 2.1 we know that the left hand side of (14) is positive for $\bar{p}>0.5$. From Result 2.2 we know that the left hand side of (14) strictly decreases in δ. To establish Result

2.3 it is thus sufficient to show that for any given $\bar{p}$ there exists one δ which turns equation (14) negative.

For this purpose, let us assume that

$$\delta = 1 - \bar{p}. \tag{16}$$

Substituting (16) into (14) and rearranging the left hand side leaves us with

$$-\left(\frac{3}{2}\cdot(1-\bar{p})\right)^2, \tag{17}$$

which is always negative. This establishes Result 2.3.

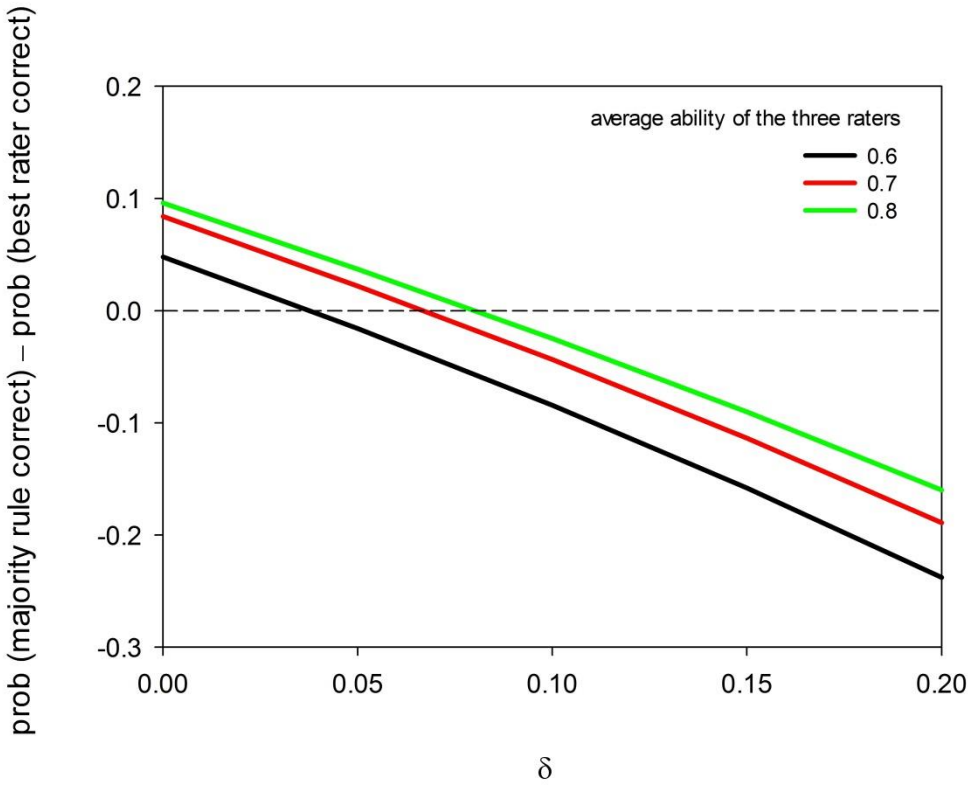

Fig. S14. As the similarity between the three diagnosticians decreases (i.e. δ increases), the probability with which these three diagnosticians – when adopting the majority rule – outperform the best diagnostician decreases. This effect is illustrated for three different average abilities $\bar{p}$ of the three diagnosticians.